Prediction of 2024 US election …[*]

Colin Sihan Yang   Lexun Yu   Siddharth Gowda

November 2, 2024

We forecast the winner of the 2024 US presidential election using “poll-of-polls” by building a linear model.

1 Introduction

Election result forecasting has become an essential tool for analysts in political science and the public to predict the outcome of democratic process, such as the presidential election in the United States. Traditionally, individual polls have been used as a snapshot of voter sentiment, but they only reflect temporary changes in the performance of contestants, instead of a precise estimation of the election result. As discussed by Pasek (2015) and Blumenthal (2014), the aggregation of multiple polls, or “poll-of-polls,” has become a popular technique to reduce individual survey errors and provide more accurate election forecasts. However, the traditional poll aggregation does not reflect dynamics of an election, especially with real-time changes and the introduction of new data. This creates a gap for a more adaptable model to predict the election result based on both polling data and additional variables, such as historical data and economic indicators.

This paper fills the gap by building a hybrid election forecasting model following the strategies mentioned by Pasek (2015). As Pasek (2015) described in their article, aggregation involves determining which surveys are worth including, as well as selecting, combining and averaging results from multiple polls to reduce individual biases and errors. Prediction modeling adds other data to the model that predicts election outcomes based on current dynamics. Hybrid models like the Bayesian approach incorporates prior beliefs based on historical data or expert knowledge and new evidence like economic updates to dynamically adjust the forecast as the campaign progresses.

In this paper, we aim to predict the 2024 us election result with the hybrid election forcasting model. We incorporate aggregation by filtering the polls on FiveThirtyEight (2024) by

[*] Code and data are available at: https://github.com/yulexun/uselection.

numeric grade that indicates pollster's reliability, prediction that incorporates social and economic indicators including unemployment rates and abortion rates, and hybrid approaches that leverages Bayesian techniques which combines historical data such as the 2016 election data, allowing for a dynamic prediction of the U.S. presidential election.

The estimand for this research paper is the predicted support percentages for Kamala Harris and Donald Trump. The prediction is based on quantifying various polling factors, including sample size, poll scores, and transparency scores, which are used as predictors.

The results of this model indicate a more stable and accurate forecast compared to traditional aggregation methods alone, [update this …]

The remainder of this paper is structured as follows: [update this …]

2 Data

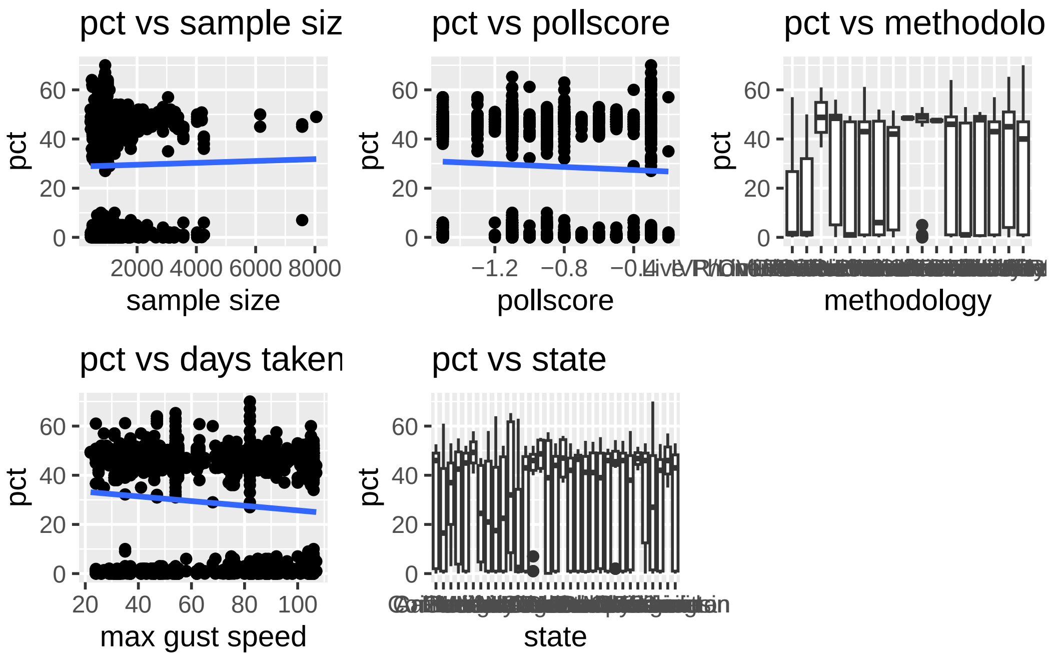
2.1 Overview

For the data we used in this analysis about the polling result for Kamala Harris and Donalad Trump in 2024 USA president election.
- **response variable:** `pct`(pct: The percentage of the vote or support that the candidate received in the poll)
- **numeric predictor:**
`sample size`(sample_size: The total number of respondents participating in the poll)
`timegap`(the time gap between the poll start date and the real election date i.e timegap = real US election date - poll start date)
`pollscore`(A numeric value representing the score or reliability of the pollster in question)
- **categorical predictor** `state`(The U.S. state where the poll was conducted or focused)
`methodology`(The method used to conduct the poll)

2.2 Data Exploration

- pct vs sample size: This scatter plot shows the pct against the sample size, with a fitted trend line indicating a slight positive relationship. The data points are denser for lower sample sizes, suggesting that smaller sample sizes are more common in the dataset.
- pct vs pollscore: This scatter plot illustrates the pct against pollscore. The fitted trend line suggests a weak negative relationship between pollscore and pct. The points are scattered without a strong linear pattern.

Table 1



- pct vs methodology: A boxplot comparing pct for different polling methodologies. The pct distribution varies across methodologies, with some showing greater spread or median differences. This suggests that the polling methodology may influence pct outcomes.
- pct vs days taken from election: A scatter plot displaying pct versus the number of days before the election. The trend line indicates a slight negative relationship, suggesting that as the election date approaches, pct may decrease slightly.
- pct vs state: A boxplot depicting pct across different states. The pct distribution varies by state, with some states showing wider variability or different median values, implying state-specific effects on pct.

Figure 1 The pairs plot displays scatter plots of four numeric variables (`pct`, `sample_size`, `pollscore`, and `days_taken_from_election`) to visualize their relationships. The data shows clustering, particularly in `pct` versus `sample_size`, suggesting potential heteroscedasticity. The `sample_size` variable is skewed towards lower values, while `pollscore` and `days_taken_from_election` have a more even spread, though `pollscore` shows central clustering. No strong linear relationships are immediately apparent between the variables, indicating that correlations are likely weak.

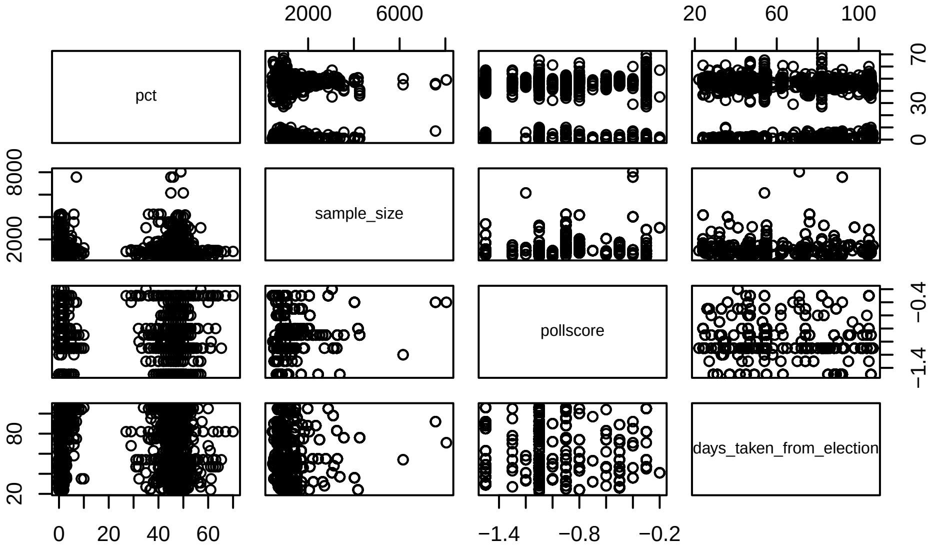


Figure 1

2.3 Measurement

In this dataset, each row represents a polling question that records the variables of interest. Each entry allows us to explore the real-world relationships between polling factors and the support percentage (`pct`) for the candidates Kamala Harris and Donald Trump. This dataset enables an analysis of how various polling characteristics influence the reported support levels for the candidates we are focused.

2.4 Clean Data

Table 2 The data cleaning process involves several steps to ensure the quality and relevance of the polling data. First, we filter the dataset to retain only poll results with a numeric grade of 2.7 or higher, indicating that the polls are considered reliable. Next, we address missing values in the state attribute: polls with NA in the state column are considered national polls.

We then create a new attribute, days_taken_from_election, which represents the time gap between the poll's start date and the actual U.S. election date. Additionally, we filter the dataset to include only polls conducted after July 21, 2024, the date when Kamala Harris declared her candidacy. Finally, we remove any remaining rows that contain missing values to ensure a clean dataset.

Table 2: Sample of cleaned US election data

pct	sample_size	pollscore	days_taken_from_election	state	methodology	candidate_name
47.6	4180	-0.8	24	National	Online Ad	Kamala Harris
50.7	4180	-0.8	24	National	Online Ad	Donald Trump
0.8	4180	-0.8	24	National	Online Ad	Jill Stein
0.1	4180	-0.8	24	National	Online Ad	Chase Oliver
0.1	4180	-0.8	24	National	Online Ad	Cornel West
48.1	4180	-0.8	24	National	Online Ad	Kamala Harris

2.5 Basic Statistics Summary for Data

Figure 2a The histogram displays the average support percentage for Kamala Harris across different U.S. states. The data indicates that support for Harris varies significantly across states. Notable observations include relatively high average support in states such as California and New York, which are known for being more Democratic-leaning. On the other hand, there are states with lower average support percentages, particularly in more traditionally Republican or swing states. The distribution suggests regional variations in support, with some states showing consistent backing for Harris while others indicate a weaker performance.

Figure 2b The second histogram shows the average support percentage for Donald Trump across various states. It highlights substantial support in states such as Florida and Texas, which align with historical trends of strong Republican support. Trump's average support appears robust in many midwestern and southern states, which are known for their conservative voter base. However, in more liberal-leaning states such as California and New York, the average support is lower, reflecting these states' tendency to lean Democratic.

3 Model

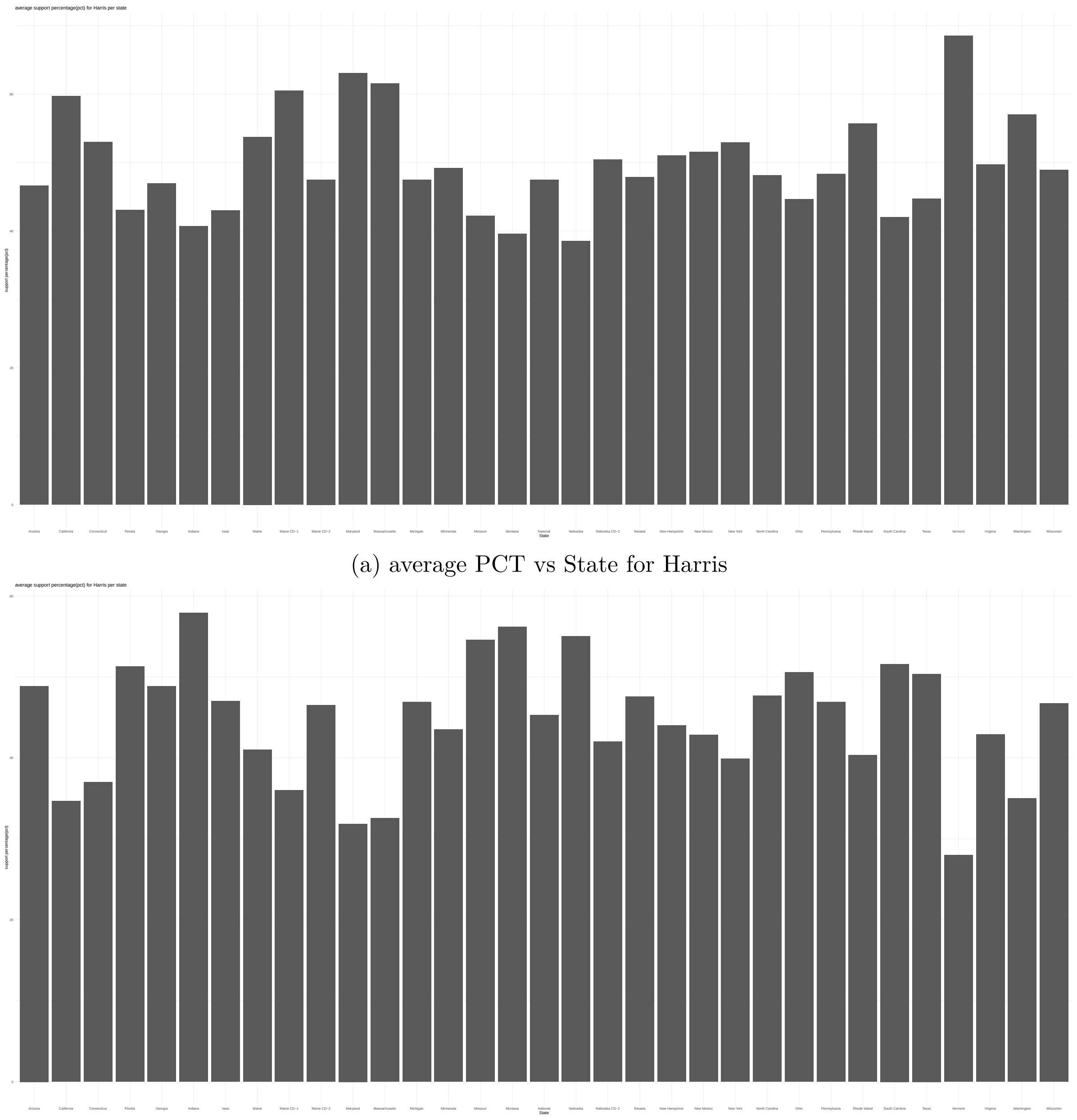
The goal of our modeling strategy is twofold. Firstly, we aim to accurately predict the support percentage (PCT) for Harris and Trump based on relevant poll data and key influencing factors. Secondly, we seek to evaluate the efficacy of different modeling approaches—from simple linear regression (SLR) to multiple linear regression (MLR) and Bayesian hierarchical models—to understand their predictive capabilities and assess the underlying relationships between variables. By comparing these models, we can determine which approach provides the most robust and reliable predictions, while considering the variability and potential uncertainty in the data.

3.1 Model set-up

The Bayesian model is implemented in R (R Core Team 2023) using the rstanarm package as described by (**rstanarm?**). The model is run with the following specifications:

- Formula: pct $\sim$ pollscore + days taken from election + sample size + (1|methodology) + (1|state)
- Priors: Normal(0, 2.5) for all coefficients and intercept, Exponential(1) for σ
- Settings: Seed = 123, Cores = 4, Adapt delta = 0.95

We run the model in R (R Core Team 2023) using the `rstanarm` package of (**rstanarm?**). We use the default priors from `rstanarm`. us



(a) average PCT vs State for Harris

(b) PCT vs State for Trump

Figure 2: the average PCT vs State for Harris and Trump

3.2 Basic Model

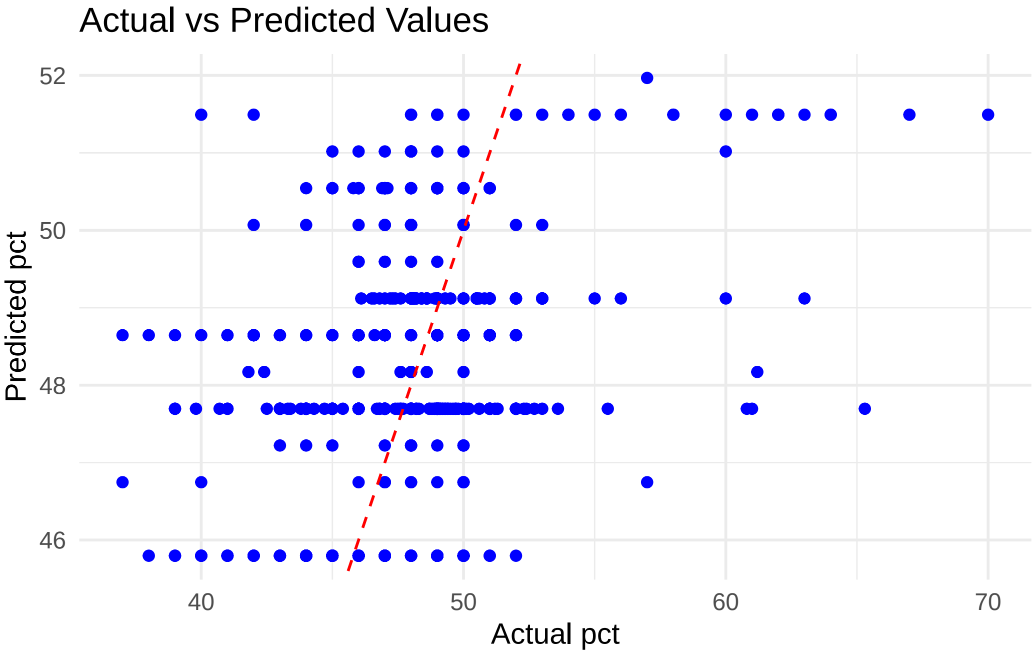


Figure 3

Figure 3 in this SLR model, the response variable is pct and the only one predictor is the pollscore. The primary concern lies in the evident dispersion of data points, which are widely spread and do not cluster closely around the line of perfect prediction (the dashed red line). This suggests that while pollscore may have some predictive capability, it does not adequately explain the variability in pct. The observed inconsistencies between actual and predicted values indicate that the relationship between pct and pollscore is likely not sufficiently captured by a linear model with just one predictor.

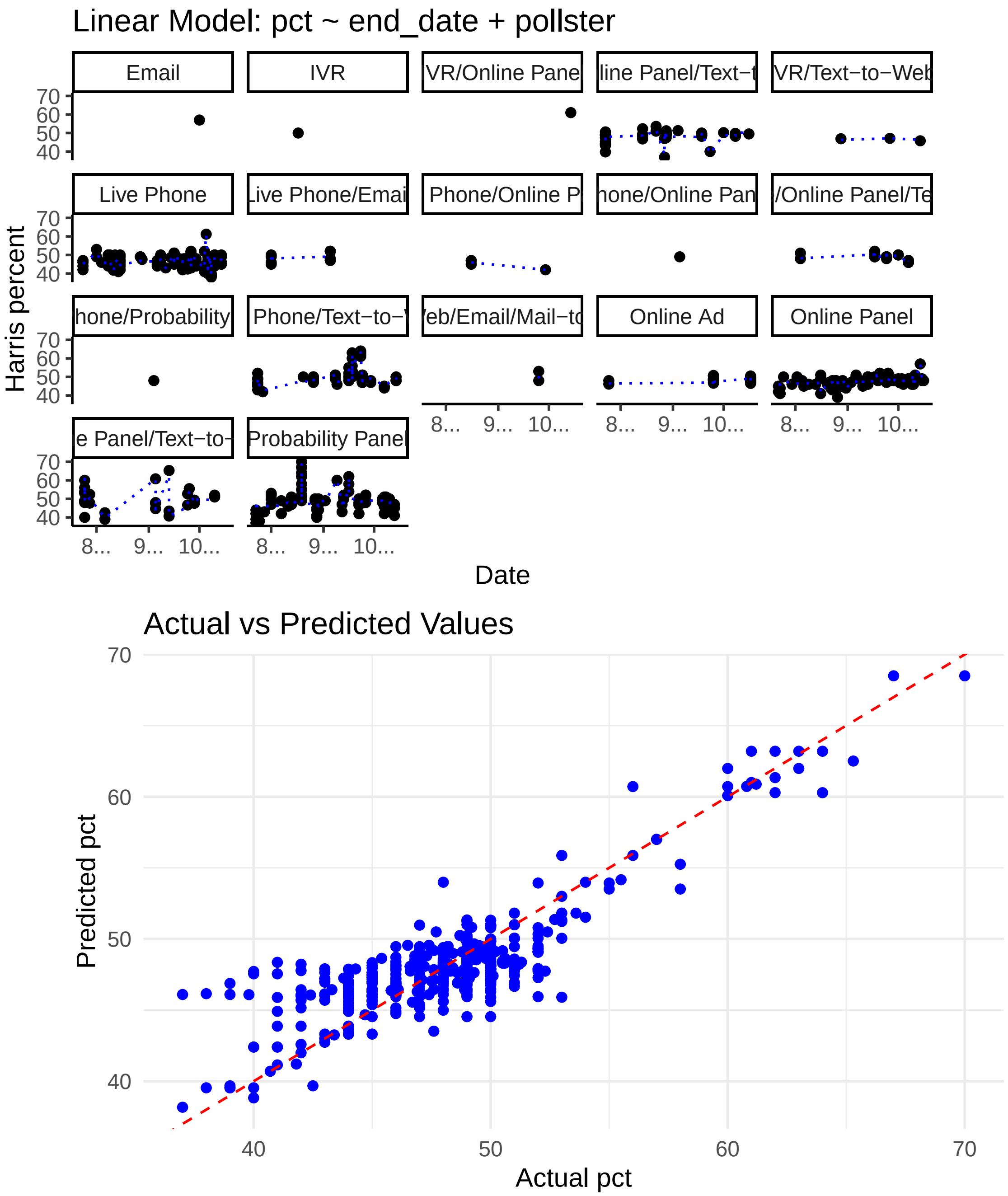


Figure 4

Figure 4, in this MLR model, we predicts pct using pollscore, days_taken_from_election, methodology, sample_size, and state as predictors. The overall distribution of points suggests that the MLR model has captured a substantial portion of the variance in pct. The majority of the data points align relatively well with the red dashed line, particularly within the middle range of pct values (approximately between 40 and 60). This alignment indicates that the model performs reasonably well in this range, with predicted values correlating strongly with the actual observed outcomes.

3.3 Bayesian Model

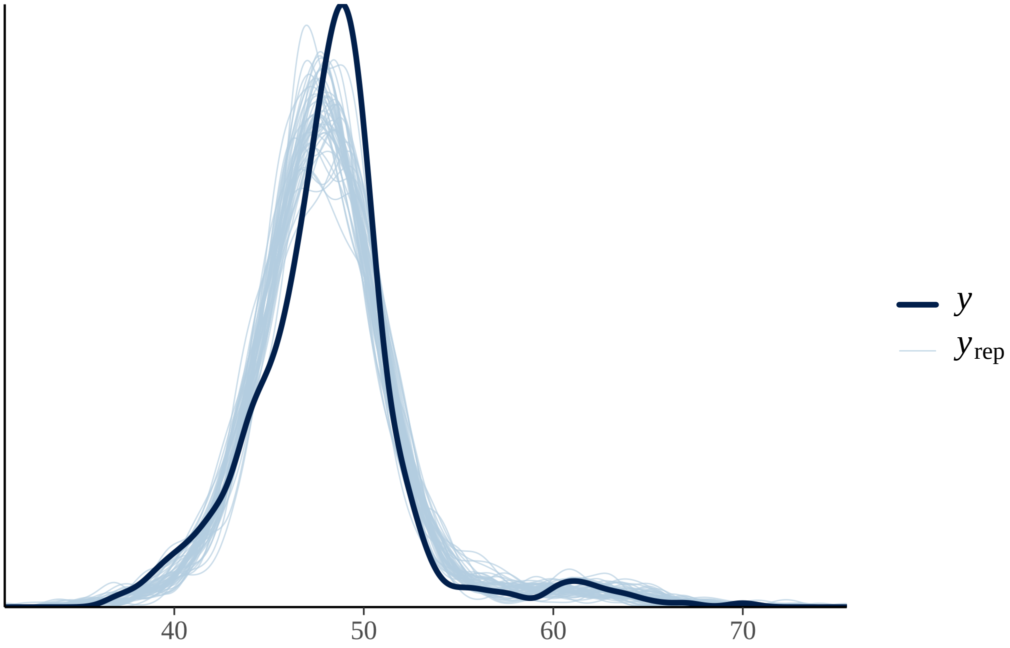


Figure 5

3.3.1 Posterior Predictive Checks

Figure 5 This plot displays the results of a posterior predictive check (PPC) for a Bayesian model using the pp_check() function.

The observed data (y) follows a distinct, symmetrical distribution centered around a specific value, suggesting a well-defined peak with tapering on both sides. The replicated distributions (y_rep), shown as multiple thin lines, generally align with the shape of the observed distribution, indicating that the Bayesian model has captured the main characteristics of the data. However, the variability in the y_rep lines highlights the degree of uncertainty inherent in the model's predictive capability.

The fact that the y_rep curves closely match the overall pattern of the actual y suggests that the model performs reasonably well in replicating the observed data. Minor discrepancies or deviations between the y and y_rep might imply areas where the model could be fine-tuned or adjusted to improve accuracy. Overall, this PPC indicates that the model provides a decent fit to the data, with some variability accounted for in the predictive samples.

3.3.2 Train Test Validation

```r
# Split data into training and test sets
set.seed(123)
train_indices <- sample(seq_len(nrow(just_harris_data)), size = 0.7 * nrow(just_harris_data)
train_data <- just_harris_data[train_indices, ]
test_data <- just_harris_data[-train_indices, ]

# Fit the model on the training data
bayesian_model_train <- stan_glmer(
  formula = formula,
  data = train_data,
  family = gaussian(),
  prior = priors,
  prior_intercept = priors,
  seed = 123,
  cores = 4,
  adapt_delta = 0.95
)

# Predict on the test data and check performance
predictions <- posterior_predict(bayesian_model_train, newdata = test_data)
```

```r
ss_total <- sum((test_data$pct - mean(test_data$pct))^2)
ss_residual <- sum((test_data$pct - colMeans(predictions))^2)
r_squared <- 1 - (ss_residual / ss_total)
print(paste("R-squared:", r_squared))
```

```
[1] "R-squared: 0.700566347978073"
```

```r
# "R-squared: 0.700563202695799"
```

$$SS_{\text{total}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{1}$$

$$SS_{\text{residual}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

$$R^2 = 1 - \frac{SS_{\text{residual}}}{SS_{\text{total}}} \tag{3}$$

$$\text{R-squared} = 0.700563202695799 \tag{4}$$

Where:

- y_i represents the actual values,
- $\bar{y}$ is the mean of the actual values,
- $\hat{y}_i$ are the predicted values,
- SS_{total} is the total sum of squares,
- SS_{residual} is the sum of squared residuals.

we could see that the R squared value equal to the 0.700563202695799 suggesting that the model captures a substantial portion of the variability in the data and demonstrates that the model performs well in explaining the variability of the pct in the test data, reflecting strong predictive capabilities.

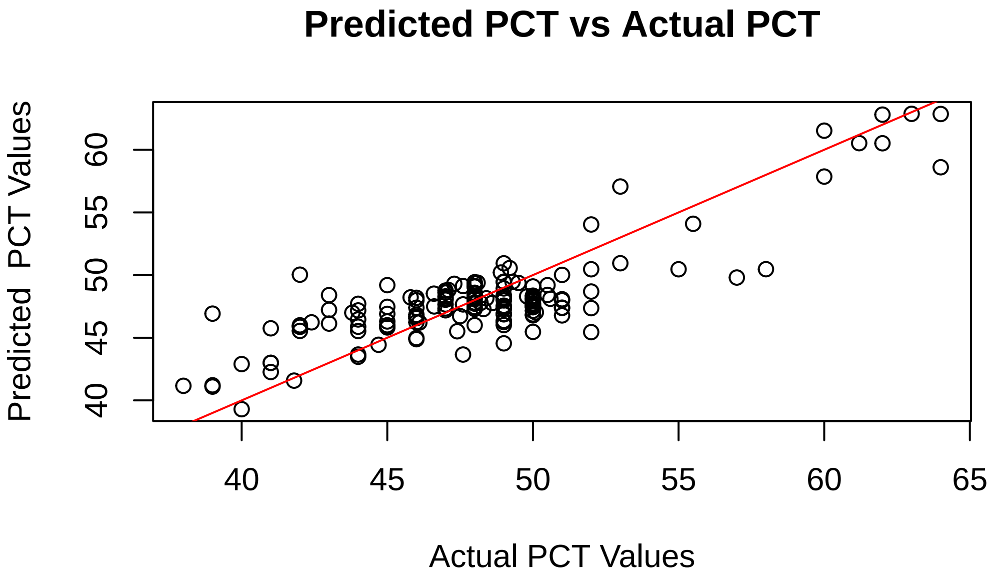


Figure 6

by visual checking actual pct and predicted pct plot Figure 6 the points are plotted against a 45-degree line, which represents the ideal scenario where predicted values match the actual values perfectly.

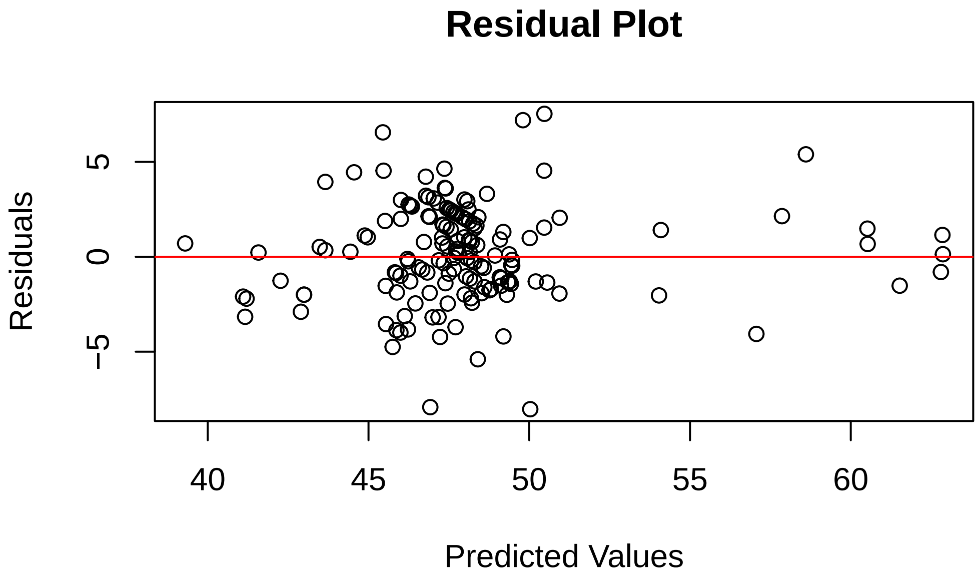


Figure 7

and the residual plot Figure 7 are fairly centered around zero with no major trend, suggesting that the model is not heavily biased in its predictions.

In conculsion, the visual checks from the predicted vs. actual plot and the residual plot, there is no strong evidence that the Bayesian model is overfitting. It appears to generalize well to the data it was trained on without showing signs of capturing noise or irrelevant patterns.

3.3.3 Model justification

To predict the support percentage (PCT) for Harris and Trump, we employed a comprehensive modeling strategy involving Simple Linear Regression (SLR), Multiple Linear Regression (MLR), and a Bayesian hierarchical model. The SLR model provided a foundational analysis of the relationship between PCT and pollscore, highlighting its limited predictive power due to a lack of complexity. The MLR model improved on this by incorporating additional predictors such as days taken from the election, sample size, methodology, and state, enhancing its predictive accuracy and capturing interactions among variables. The Bayesian hierarchical model further refined our approach, incorporating prior knowledge and accounting for variability at group levels (e.g., methodology and state) through random intercepts. This model provided robust uncertainty quantification through credible intervals and demonstrated strong predictive performance with an R-squared value around 0.70, indicating it effectively explained data variability. By leveraging the strengths of these models, particularly the Bayesian approach's interpretability and robustness, we achieved reliable predictions and a deeper understanding of the factors influencing support percentages.

4 Result

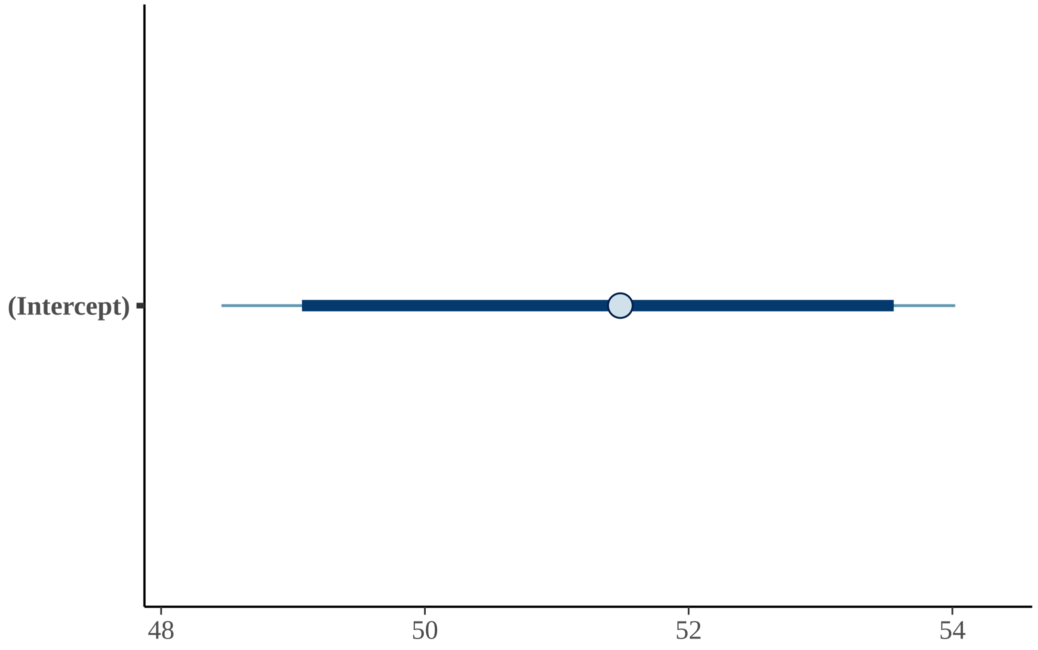


Figure 8

Figure 8 This plot displays the estimated posterior distribution for the intercept parameter in the Bayesian model. The intercept represents the baseline level of support, expressed as a percentage, for Kamala Harris in the U.S. election data used in the analysis. The plot features a point estimate (depicted by the central dot) that indicates the mean or median of the posterior distribution for the intercept, and a horizontal line showing the 95% credible interval, which signifies the range within which the true value of the intercept is likely to fall with 95% probability.

The range of this credible interval spans from approximately 50 to 54 percent, suggesting that the model estimates the baseline level of support for Kamala Harris to be around this range. The relatively narrow width of the interval implies a certain degree of confidence in the estimate, indicating that the data used in the model provided a clear signal for the intercept's value.

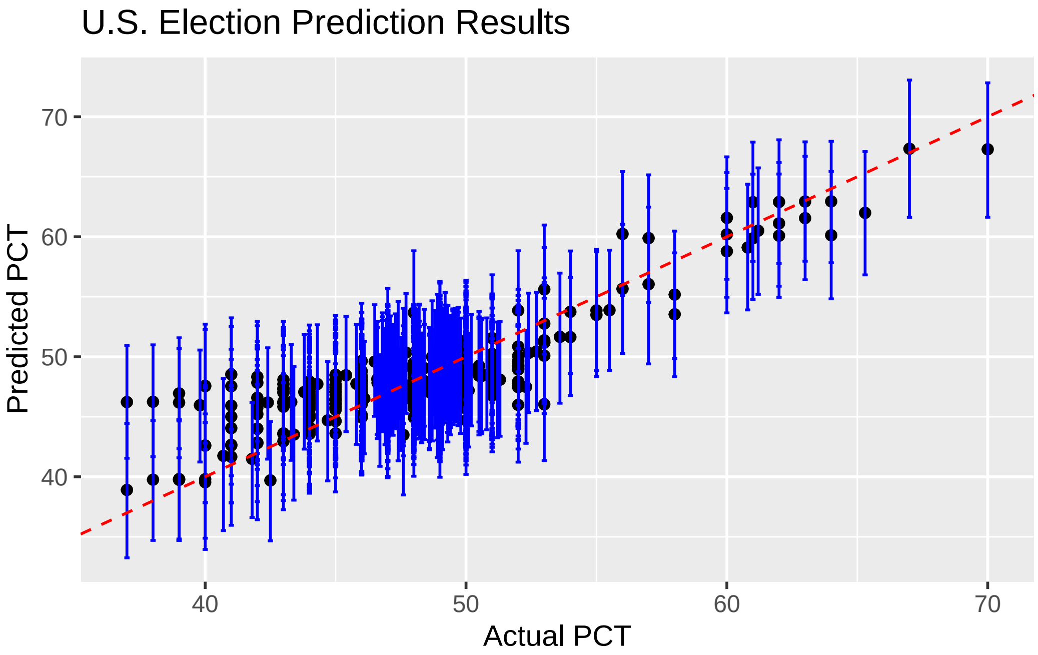
In conclusion, the estimated intercept, which represents the baseline percentage of support for Kamala Harris in the U.S. election data analyzed, is centered around 52%, with a credible interval spanning approximately from 50% to 54%. This suggests that, according to the Bayesian model, the underlying support for Kamala Harris.

```r
predictions2 <- posterior_predict(bayesian_model_1, newdata = just_harris_data)

predicted_means <- colMeans(predictions2)
predicted_intervals <- apply(predictions2, 2, quantile, probs = c(0.025, 0.975))

# Combine the results with the actual data
result_summary <- data.frame(
  Actual_PCT = just_harris_data$pct,
  Predicted_PCT = predicted_means,
  Lower_CI = predicted_intervals[1, ],
  Upper_CI = predicted_intervals[2, ]
)

ggplot(result_summary, aes(x = Actual_PCT, y = Predicted_PCT)) +
  geom_point() +
  geom_errorbar(aes(ymin = Lower_CI, ymax = Upper_CI), width = 0.2, color = "blue") +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  labs(
    title = "U.S. Election Prediction Results",
    x = "Actual PCT",
    y = "Predicted PCT"
  )
```



```
print(head(result_summary))
```

```
  Actual_PCT Predicted_PCT Lower_CI Upper_CI
1       47.6      49.12850 44.44701 54.04904
2       48.1      49.13812 44.34297 53.83691
3       48.6      47.74757 42.95339 52.43367
4       49.3      47.71647 42.90265 52.32287
5       48.1      48.02783 43.50669 52.78940
6       48.4      47.98070 43.11644 52.70039
```

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows... This suggests...

?@fig-stanareyouokay-2 is a Rhat plot. It shows... This suggests...

C FiveThirtyEight Licenses

FiveThirtyEight's data sets are used and modified by us under the Creative Commons Attribution 4.0 International License.

References

Blumenthal, Mark. 2014. "Polls, Forecasts, and Aggregators." *PS: Political Science and Politics* 47 (2): 297–300. http://www.jstor.org/stable/43284537.

FiveThirtyEight. 2024. "Our Data." *FiveThirtyEight*. https://data.fivethirtyeight.com.

Pasek, Josh. 2015. "THE POLLS–REVIEW: PREDICTING ELECTIONS: CONSIDERING TOOLS TO POOL THE POLLS." *The Public Opinion Quarterly* 79 (2): 594–619. http://www.jstor.org/stable/24546379.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.